A PROOFS FOR ANALYSIS OF GRADIENT DESCENT

In order to highlight our main contribution conceptually, we simplified the statements of the theorems and lemmas stated in the main body for exposition. Hence, in this section, we shall restate Theorem 1.1 and the key lemmas formally and present the complete proof for our main theorem. The formal statement of the main theorem is as follows

Theorem 1.1 (Formal version of the main theorem) *For any absolute constants $C_1 \geq 1, C_2 > 0$, there exists absolute constants $c_3 > 0, c_\eta > 0$ such that the following holds. Suppose $\widetilde{\mathcal{D}}_x$ be a distribution over $(\widetilde{x}, y) \in \mathbb{R}^d \times \mathbb{R}$ where the marginal over $\widetilde{x}$ is the standard Gaussian $\mathcal{N}(0, I)$, $H := \{(\widetilde{w}, b_w) : \|\widetilde{w}\| \in [1/C_1, C_1], |b_w| \leq C_2\}$, and consider population gradient descent iterates $w_{t+1} = w_t - \eta \nabla L(w)$, with the initializer $w_0 = (\widetilde{w}_0, 0)$ where $\widetilde{w}_0$ is drawn from the radially symmetric distribution in Section 5. For any $\varepsilon > 0$ and learning rate $\eta = c_\eta d^{-1}$, with at least constant probability $c_3 > 0$, one of the iterates w_T of population gradient descent after $\mathrm{poly}(d, 1/\varepsilon)$ steps satisfies $L(w_T) = O(OPT) + \varepsilon$.*

Note that without loss of generality, we can assume $\varepsilon \leq O(OPT)$. If we cannot make such assumption (e.g. when $OPT \approx 0$), we can use an upper bound on OPT of $O(\varepsilon)$, and carry out the same analysis.

Recall that $v = (\tilde{v}, b_v) \in \mathbb{R}^{d+1}$ is any minimizer of the loss $L(w)$ i.e., $v := \arg\min_{w \in H} L(w)$. Hence $L(v) = OPT$. We will assume in the rest of the analysis that $\|\tilde{v}\|_2 = 1$ for simplifying our exposition. But this is not necessary; when $\|\tilde{v}\|_2 \in [1/C_1, C_1]$, we can carry out the same analysis incurring some extra constant factors. Finally recall that the realizable loss

$$F(w) := \frac{1}{2} \mathbb{E}\left[(\sigma(w^\top x) - \sigma(v^\top x))^2\right].$$

Our goal is to prove that for some iterate T, we have $F(w_T) \leq O(OPT)$. This implies that $L(w_T) \leq 2F(w_T) + 2OPT = O(OPT)$.

To prove Theorem 1.1, we first formally establish the following useful lemmas.

Lemma 4.1 (Lower bound on the measure of the intersection)**.** *Suppose the marginal distribution $\widetilde{\mathcal{D}}_x$ over $\widetilde{x}$ is $O(1)$-regular. There exists an absolute constant $c > 0$ such that for all $\delta > 0$, if $F(w) \leq F(0) - \delta$ then*

$$\mathbb{P}[w^\top x \geq 0, v^\top x \geq 0] \geq \frac{\delta^2}{c\|w\|_2^4 \|v\|_2^4} = \frac{\delta^2}{c\|w\|_2^4 (1 + |b_v|^2)^2}. \tag{8}$$

Proof. Recall that $F(x)$ is the realizable loss i.e., the loss compared to the optimal solution v. Since $F(w) \leq F(0) - \delta$, we have

$$\begin{aligned} F(0) - \delta \geq F(w) &:= \frac{1}{2} \mathbb{E}[(\sigma(w^\top x) - \sigma(v^\top x))^2] \\ &= \frac{1}{2} \mathbb{E}[\sigma(w^\top x)^2] - \mathbb{E}[\sigma(w^\top x)\sigma(v^\top x)] + \frac{1}{2} \mathbb{E}[\sigma(v^\top x)^2] \\ &\geq - \mathbb{E}[\sigma(w^\top x)\sigma(v^\top x)] + F(0) \\ \text{Hence } \delta &\leq \mathbb{E}[\sigma(w^\top x)\sigma(v^\top x)]. \end{aligned} \tag{12}$$

Moreover we can also get an upper bound on $\mathbb{E}[\sigma(w^\top x)\sigma(v^\top x)]$ using Cauchy-Schwartz inequality and repeated applications of Young's inequality.

$$\begin{aligned}
\mathbb{E}[\sigma(w^\top x)\cdot\sigma(v^\top x)] &= \mathbb{E}\left[\mathbf{1}[w^\top x\geq 0, v^\top x\geq 0](w^\top x)(v^\top x)\right]\\
&\leq \sqrt{\mathbb{P}[w^\top x\geq 0, v^\top x\geq 0]}\cdot\sqrt{\mathbb{E}[(w^\top x)^2(v^\top x)^2]}\\
&\leq \sqrt{\mathbb{P}[w^\top x\geq 0, v^\top x\geq 0]}\cdot\sqrt{2\,\mathbb{E}[(w^\top x)^4]+2\,\mathbb{E}[(v^\top x)^4]}\\
&\leq 8\sqrt{\mathbb{P}[w^\top x\geq 0, v^\top x\geq 0]}\cdot\sqrt{\mathop{\mathbb{E}}_{\widetilde{x}\sim N(0,I)}[(\tilde{w}^\top\widetilde{x})^4]+b_w^4}\cdot\sqrt{\mathop{\mathbb{E}}_{\widetilde{x}\sim N(0,I)}[(\tilde{v}^\top\widetilde{x})^4]+b_v^4}\\
&\leq 8\sqrt{\mathbb{P}[w^\top x\geq 0, v^\top x\geq 0]}\cdot\sqrt{(\beta_4\|\tilde{w}\|_2^4+b_w^4)\cdot(\beta_4\|\tilde{v}\|_2^4+b_v^4)}\\
&\leq c'\sqrt{\mathbb{P}[w^\top x\geq 0, v^\top x\geq 0]}\cdot\|w\|_2^2\|v\|_2^2. \qquad (13)
\end{aligned}$$

for some constant $c'>0$, where the last but one line follows from the standard bounds on the fourth-moment of an $O(1)$-regular distribution. Combining (12) and (13) concludes the lemma.

□

Lemma 4.2 (Improvement from the first order term) *Suppose the marginal over $\widetilde{x}$ is $O(1)$-regular. There exists absolute constant $c_1>0$ such that for any $\delta>0$, if $\|v\|_2,\|w\|_2\leq B$ and $F(w)\leq F(0)-\delta$ then*

$$\langle\nabla F(w), w-v\rangle\geq\gamma\|w-v\|^2,\ \textit{where}\ \gamma=\frac{c_1\delta^9}{B^{28}}. \qquad (14)$$

The constant c_1 depend on the constants $\beta_1,\beta_2',\beta_2,\beta_4$ etc. in the regularity assumption of $\widetilde{\mathcal{D}}_x$.

We remark that for our setting of parameters $\delta=\Omega(1)$ and $B=O(1)$, and hence we will conclude that $\langle\nabla F, w-v\rangle\geq\Omega(\|w-v\|_2^2)$.

Proof. This lemma only concerns the "realizable portion" of the loss function $F(w)$.

Let $u=(\tilde{u},b_u)\in\mathbb{R}^{d+1}$ be the unit vector along $w-v$. We have

$$\begin{aligned}
\langle\nabla F(w), w-v\rangle &= \mathbb{E}\left[\big(\sigma(w^\top x)-\sigma(v^\top x)\big)\sigma'(w^\top x)(w^\top x-v^\top x)\right]\\
&= \mathbb{E}\left[\big(w^\top x-v^\top x\big)^2\mathbf{1}[w^\top x\geq 0, v^\top x\geq 0]\right]\\
&\quad+\mathbb{E}\left[w^\top x\big(w^\top x-v^\top x\big)\mathbf{1}[w^\top x\geq 0, v^\top x<0]\right]\\
&\geq \mathbb{E}\left[\big(w^\top x-v^\top x\big)^2\mathbf{1}[w^\top x\geq 0, v^\top x\geq 0]\right]\\
&= \|w-v\|_2^2\cdot\mathbb{E}\left[(u^\top x)^2\mathbf{1}[w^\top x\geq 0, v^\top x\geq 0]\right] \qquad (15)
\end{aligned}$$

Let $q:=c\delta^2/B^8$ and $\tau:=c'\frac{\delta^4}{B^{16}}$ for some sufficiently small absolute constant $c'>0$ that will be chosen later. We will now lower bound the contribution from just the samples that achieve a value $(u^\top x)^2>\tau^2$ using Lemma 4.1, that lower bounds $\mathbb{P}[w^\top x\geq 0, v^\top x\geq 0]\geq c\delta^2/B^8=q$:

$$\begin{aligned}
&\mathbb{E}\left[(u^\top x)^2\mathbf{1}[w^\top x\geq 0, v^\top x\geq 0]\right]\\
&\geq\tau^2\cdot\mathop{\mathbb{P}}_x\left[w^\top x\geq 0, v^\top x\geq 0, (u^\top x)^2\geq\tau^2\right]\\
&\geq\tau^2\cdot\left(\mathop{\mathbb{P}}_x\left[w^\top x\geq 0, v^\top x\geq 0\right]-\mathop{\mathbb{P}}_x\left[(u^\top x)^2<\tau^2\right]\right)\\
&\geq\tau^2\cdot\left(q-\mathop{\mathbb{P}}_{\widetilde{x}}[|\tilde{u}^\top x+b_u|<\tau]\right). \qquad (16)
\end{aligned}$$

Now we just need to upper bound $\mathbb{P}_{\widetilde{x}}[|\tilde{u}^\top\widetilde{x}+b_u|<\tau]$. Let $\beta=\|\tilde{u}\|_2$. If $\beta=\|\tilde{u}\|_2\ll|b_u|$, then $|b_u|$ is itself large, and $\tilde{u}^\top\widetilde{x}$ is too small in comparison to bring down $|\tilde{u}^\top\widetilde{x}+b_u|$. On the other hand,

if $\beta = \|\tilde{u}\|_2$ is not too small, then the anti-concentration (or spread out density) of the distribution $\widetilde{\mathcal{D}}_x$ ensures that $|\tilde{u}^\top \widetilde{x} + b_u|$ is small with very low probability. We now formalize this intuition.

Suppose $\beta = \|\tilde{u}\|_2 \leq \frac{1}{4\beta_4}(q/2)^{1/4}$. Then $|b_u| > 1/2$, since $\|u\|_2 = 1$. Also from our choice, $\tau < 1/4$. Hence by the bounded fourth moments property of $\widetilde{\mathcal{D}}_x$ and Markov's inequality,

$$\mathop{\mathbb{P}}_{\widetilde{x}\sim\widetilde{\mathcal{D}}_x}[|\tilde{u}^\top \widetilde{x} + b_u| < \tau] \leq \mathop{\mathbb{P}}_{\widetilde{x}\sim\widetilde{\mathcal{D}}_x}[|\tilde{u}^\top \widetilde{x}| > \tfrac{1}{4}] \leq \frac{\mathbb{E}_{\widetilde{x}\sim\widetilde{\mathcal{D}}_x}[\langle \tilde{u}, \widetilde{x}\rangle^4]}{(1/4)^4} \leq \beta_4(4\|\tilde{u}\|_2)^4 \leq \frac{q}{2}.$$

On the other hand, if $\beta = \|\tilde{u}\|_2 > \frac{1}{4\beta_4}(q/2)^{1/4}$. Suppose $\hat{u}$ is the unit vector along $\tilde{u}$. Then using the fact that $\hat{u}^\top \widetilde{x} = \tilde{u}^\top \widetilde{x}/\|\tilde{u}\|$ is anti-concentrated by the properties of $\widetilde{\mathcal{D}}_x$. Hence we have for some constant $\beta_3 > 0$

$$\mathop{\mathbb{P}}_{\widetilde{x}\sim\widetilde{\mathcal{D}}_x}\Big[|\tilde{u}^\top \widetilde{x} + b_u| < \tau\Big] = \mathop{\mathbb{P}}_{\widetilde{x}\sim\widetilde{\mathcal{D}}_x}\Big[\hat{u}^\top \widetilde{x} \in (\tfrac{b_u-\tau}{\|\tilde{u}\|} - \tfrac{b_u+\tau}{\|\tilde{u}\|})\Big] \leq \frac{\beta_3\tau}{\|\tilde{u}\|} \leq 32\beta_3\beta_4\Big(\frac{2}{q}\Big)^{1/4}\tau < \frac{q}{2},$$

from our choice of parameters since $\tau = c'\delta^{5/2}/(B^{10}\beta_3\beta_4)$ for a sufficiently small $c' > 0$. Substituting back in (16) and (15) we have

$$\langle \nabla F(w), w - v\rangle \geq \tau^2 \cdot \frac{q}{2} \geq c_1\|w - v\|_2^2 \cdot \frac{\delta^5}{B^{20}\beta_3^2\beta_4^2} \cdot \frac{\delta^2}{B^8} \geq c_1\|w - v\|_2^2 \cdot \frac{\delta^9}{B^{28}}.$$

□

Lemma 4.3 (Success if $\|\nabla F\| \leq O(\sqrt{OPT + \varepsilon})$) *Suppose $B, \delta > 0$ are constants such that $\|v\|_2, \|w\|_2 \leq B$ and $F(w) \leq F(0) - \delta$, and $\widetilde{x}$ follows a $O(1)$-regular distribution. Then there exists a constant $C_G > 0$, such that if $\|\nabla F(w)\| \leq C_G\sqrt{OPT + \varepsilon}$ for some $\varepsilon > 0$, then $\|w - v\|_2 \leq O(\sqrt{OPT + \varepsilon})$.*

Proof. We can first apply Lemma 4.2 to conclude that

$$\langle \nabla F(w), w - v\rangle \geq \gamma\|w - v\|^2,$$

for some constant $\gamma > 0$ (since $B, \delta > 0$ are constants). Hence

$$\|\nabla F(w)\|\|w - v\| \geq \langle \nabla F(w), w - v\rangle \geq \gamma\|w - v\|^2,$$

Thus $\|w - v\|_2 \leq O(\sqrt{OPT + \varepsilon})$ which implies the lemma. □

Lemma 4.4 (Small $\|w_t - v\|$ implies small $F(w_t)$) *If $\|w_t - v\|_2 \leq O(\sqrt{OPT + \varepsilon})$ for some $\varepsilon > 0$, then $F(w_t) \leq O(OPT + \varepsilon)$.*

Proof. Since ReLU function is 1-Lipschitz (i.e. $|\sigma(z) - \sigma(z')| \leq |z - z'|$),

$$F(w_t) = \frac{1}{2}\,\mathbb{E}\Big[(\sigma(w_t^\top x) - \sigma(v^\top x))^2\Big] \leq \frac{1}{2}\,\mathbb{E}\Big[(w_t^\top x - v^\top x)^2\Big] = \frac{\|w_t - v\|^2}{2}\,\mathbb{E}\Big[(u^\top x)^2\Big]$$

where we defined $u = \frac{w_t - v}{\|w_t - v\|}$, hence the last equation. Now, notice by using Young's inequality, we get

$$\mathbb{E}\Big[(u^\top x)^2\Big] = \mathbb{E}\Big[(\widetilde{u}^\top \widetilde{x} + b_u)^2\Big] \leq 2\,\mathbb{E}\Big[(\widetilde{u}^\top \widetilde{x})^2\Big] + 2b_u^2 = 2\|\widetilde{u}\|^2 + 2b_u^2 = 2$$

due to standard properties of Gaussian distributions. Hence

$$F(w_t) \leq \frac{\|w_t - v\|^2}{2} \cdot 2 = \|w_t - v\|_2^2 \leq O(OPT + \varepsilon)$$

which concludes the proof. □

At a high level, we follow a similar approach as in Frei et al. (2020). We aim to show that for every t, either (a) $\|w_t - v\|^2 - \|w_{t+1} - v\|^2 \geq \eta C'(OPT + \varepsilon)$ is true for some $\eta > 0$ we will specify later, or (b) $\|w_t - v\|^2 \leq O(\gamma^{-1}(OPT + \varepsilon))$ holds. Since when $\|w_t - v\|^2 \leq O(OPT)$, Lemma 4.4 indicates that $F(w_t)$ is $O(OPT)$, hence $L(w_t)$ is also $O(OPT)$; this gives the required iterate w_T of gradient descent to complete the theorem when (b) holds. Hence we shall assume at time t (b) does not hold yet, and it suffices showing (a) is true. In the next lemma, we argue that throughout gradient descent, the distance between the current iterate w_t and the target weight v, $\|w_t - v\|^2$ continues to decrease as long as w_t is not too close to v.

Lemma 4.5 (Decrease in $\|w_t - v\|$: formal version) *Assume at time t, $F(w_t) \le F(0) - \delta$ where $\delta > 0$ is a constant and $\widetilde{\mathcal{D}}_x$ is $O(1)$-regular. For constants $\eta = \frac{0.05\cdot\gamma}{d\beta_2}, C_p = \frac{1}{9}(\sqrt{\frac{100\beta_2^2/\gamma^2+90}{\beta_2/\gamma}} + 10\sqrt{\frac{\beta_2}{\gamma}}), C' = 19.8\gamma/\beta_2$ where γ is defined as in Lemma 4.2, if for some $\varepsilon > 0$ $\|w_t - v\|^2 > \gamma^{-1}C_p^2(OPT + \varepsilon)$, then $\|w_{t+1} - v\|^2 \le \|w_t - v\|^2 - \eta C'(OPT + \varepsilon)$.*

Proof. Note at each timestep t,

$$w_{t+1} = w_t - \eta\nabla L(w_t) \tag{17}$$

$$w_{t+1} - v = w_t - v - \eta\nabla L(w_t) \tag{18}$$

$$\implies \|w_t - v\|^2 - \|w_{t+1} - v\|^2 = 2\eta\langle\nabla L(w_t), w_t - v\rangle - \eta^2\|\nabla L(w_t)\|^2 \tag{19}$$

therefore to lower-bound $\|w_t - v\|^2 - \|w_{t+1} - v\|^2$, we will give a lower bound for $\langle\nabla L(w_t), w_t - v\rangle$ and an upper bound for $\|\nabla L(w_t)\|^2$.

Lower bounding $\langle\nabla L(w_t), w_t - v\rangle$: Recall that $\nabla L(w_t) = \nabla F(w_t) + \mathbb{E}[(\sigma(v^\top x) - y)\sigma'(w_t^\top x)x]$, implying $\langle\nabla L(w_t), w_t - v\rangle = \langle\nabla F(w_t), w_t - v\rangle + \langle\mathbb{E}[(\sigma(v^\top x) - y)\sigma'(w_t^\top x)x], w_t - v\rangle$.

Since a direct application of Lemma 4.2 already gives a lower bound on $\langle\nabla F(w_t), w_t - v\rangle$, we need only focus on upper bounding $\left|\langle\mathbb{E}[(\sigma(v^\top x) - y)\sigma'(w_t^\top x)x], w_t - v\rangle\right|$. By Cauchy–Schwarz and Young's inequality, we get

$$\begin{aligned}
&\langle\mathbb{E}[(\sigma(v^\top x) - y)\sigma'(w_t^\top x)x], w_t - v\rangle = \mathbb{E}[(\sigma(v^\top x) - y)\sigma'(w^\top x)(w_t^\top x - v^\top x)] \\
&\ge -\sqrt{\mathbb{E}[(\sigma(v^\top x) - y)^2]} \cdot \sqrt{\mathbb{E}[(w_t^\top x - v^\top x)^2\sigma'(w_t^\top x)]} \\
&\ge -\sqrt{2OPT}\sqrt{\mathbb{E}[((w_t - v)^\top x)^2]} \\
&\ge -\sqrt{2OPT} \cdot \sqrt{2\beta_2}\|w_t - v\| = -2\sqrt{\beta_2} \cdot \sqrt{OPT} \cdot \|w_t - v\|
\end{aligned}$$

Putting these bounds together we get

$$\nabla L(w_t) = \nabla F(w_t) + \mathbb{E}[(\sigma(v^\top x) - y)\sigma'(w_t^\top x)x] \ge \gamma\|w_t - v\|_2^2 - 2\sqrt{\beta_2}\sqrt{OPT} \cdot \|w_t - v\|_2$$

Upper bounding $\|\nabla L(w_t)\|^2$: Define

$$H(w_t) = \mathbb{E}[\sigma(v^\top x - y)\sigma'(w^\top x)x]$$

and observe that

$$\nabla L(w_t) = \nabla F(w_t) + H(w_t)$$

For the first term,

$$\|\nabla F(w_t)\| \le \mathbb{E}[|\sigma(w_t^\top x) - \sigma(v^\top x)| \cdot |\sigma'(w_t^\top x)| \cdot \|x\|] \le \mathbb{E}[|w_t^\top x - v^\top x| \cdot \|x\|]$$

since $\sigma(\cdot)$ is 1-Lipschitz (i.e. $|\sigma(z) - \sigma(z')| \le |z - z'|$) and $\sigma'(\cdot) \le 1$. Hence, applying Cauchy-Schwarz yields

$$\le \sqrt{\mathbb{E}[|w_t^\top x - v^\top x|^2] \cdot \mathbb{E}[\|x\|^2]} \le \|w_t - v\| \cdot \sqrt{\beta_2}\sqrt{d+1}$$

Similarly, for the second term,

$$\|H(w_t)\| \le \mathbb{E}[|\sigma(v^\top x) - y| \cdot \|x\|] \le \sqrt{\mathbb{E}[|\sigma(v^\top x) - y|^2] \cdot \mathbb{E}[\|x\|^2]} \le \sqrt{2OPT} \cdot \sqrt{d+1}$$

Using the above two expression, we can hence bound $\|\nabla L(w_t)\|^2$ as

$$\|\nabla L(w_t)\|^2 \le 2\|\nabla F(w_t)\|^2 + 2\|H(w_t)\|^2 \le 4d\beta_2\|w_t - v\|^2 + 4dOPT \tag{20}$$

Lower bounding $\|w_t - v\|^2 - \|w_{t+1} - v\|^2$: The above inequalities yield

$$\begin{aligned}\|w_t - v\|^2 - \|w_{t+1} - v\|^2 &= 2\eta\langle \nabla L(w_t), w_t - v\rangle - \eta^2\|\nabla L(w_t)\|^2 \\ &\geq 2\eta \cdot \Big[\gamma\|w_t - v\|^2 - 2\sqrt{\beta_2}\sqrt{OPT}\|w_t - v\|\Big] - 4d\eta^2 \cdot (\beta_2\|w_t - v\|^2 + OPT) \\ &\geq 2\eta \cdot \Big[\gamma\|w_t - v\|^2 - 2(\beta_2\gamma)^{1/2}C_p^{-1}\|w_t - v\|^2\Big] - 4d\eta^2 \cdot (\beta_2\|w_t - v\|^2 + OPT) \\ &= 2\eta\Big(\gamma - 2(\beta_2\gamma)^{1/2}C_p^{-1} - 2d\eta\beta_2\Big)\|w_t - v\|^2 - 2\eta \cdot 2d\eta OPT\end{aligned}$$

due to our assumption that (b) does not hold yet, i.e. $\|w_t - v\| > C_p\gamma^{-1/2}\sqrt{(OPT + \varepsilon)} > C_p\gamma^{-1/2}\sqrt{OPT}$ with some constant $C_p > 0$, implying $\sqrt{OPT} < \gamma^{1/2}C_p^{-1}\|w_t - v\|$. Consequently, by choosing $\eta = \frac{0.05\cdot\gamma}{d\beta_2} = O(d^{-1})$ and $C_p = \frac{1}{9}(\sqrt{\frac{100\beta_2^2/\gamma^2+90}{\beta_2/\gamma}} + 10\sqrt{\frac{\beta_2}{\gamma}}) = O(1)$, we get

$$\begin{aligned}&\geq 2\eta\Big(100 \cdot 2d\eta \cdot OPT + 100 \cdot 2d\eta \cdot \varepsilon - 2d\eta \cdot OPT\Big) = 2\eta\Big(198d\eta \cdot OPT + 200d\eta \cdot \varepsilon\Big) \\ &\geq \eta \cdot 396d \cdot \frac{0.05 \cdot \gamma}{d\beta_2} \cdot (OPT + \varepsilon) = \eta C'(OPT + \varepsilon)\end{aligned}$$

by setting $C' = 19.8\gamma/\beta_2$. Hence the proof follows. □

We will also use as a black box two lemmas given in Vardi et al. (2021) that uses the smoothness of the function to upper bound the contribution from the second order term.

Lemma A.1 (Lemma D.4 in Vardi et al. (2021)). *For any $w, w' \in \mathbb{R}^{d+1}$, if $\widetilde{x}$ follows a $O(1)$-regular distribution and $\forall\lambda \in [0,1]$ there exists constants $C_\ell, C_u > 0$ such that $\|(1-\lambda)w + \lambda w'\| \in [C_\ell, C_u]$, then $\|\nabla F(w) - \nabla F(w')\| \leq (c_1' + \frac{8\beta_3(C_u + \sqrt{C_1^2 + C_2^2})c_2'}{C_\ell}) \cdot \|w - w'\|$ where $c_1', c_2' > 0$ are absolute constants.*

Proof. Note that the original proof of this lemma relies on the assumption that the distribution of $\widetilde{x}$ is compactly supported. Hence we shall provide a modified proof that generalizes the lemma statement to $O(1)$-regular distributions. Similar to the argument in Vardi et al. (2021), we write $\|\nabla F(w) - \nabla F(w')\|$ as

$$\begin{aligned}\|\nabla F(w) - \nabla F(w')\| &= \Big\|\mathbb{E}\Big[(\sigma(w^\top x) - \sigma(v^\top x))\sigma'(w^\top x)x\Big] - \mathbb{E}\Big[(\sigma(w'^\top x) - \sigma(v^\top x))\sigma'(w'^\top x)x\Big]\Big\| \\ &\leq \Big\|\mathbb{E}\Big[\mathbb{1}\{w^\top x \geq 0, w'^\top x \geq 0\}((w - w')^\top x)x\Big]\Big\| \\ &+ \Big\|\mathbb{E}\Big[\mathbb{1}\{w^\top x \geq 0, w'^\top x < 0\}(w^\top x - \sigma(v^\top x))x\Big]\Big\| \\ &+ \Big\|\mathbb{E}\Big[\mathbb{1}\{w^\top x < 0, w'^\top x \geq 0\}(w'^\top x - \sigma(v^\top x))x\Big]\Big\| \\ &\leq \mathbb{E}\Big[\|((w - w')^\top x)x\|\Big] \\ &+ \mathbb{E}\Big[\mathbb{1}\{w^\top x \geq 0, w'^\top x < 0\}\|(w^\top x - \sigma(v^\top x))x\|\Big] \\ &+ \mathbb{E}\Big[\mathbb{1}\{w^\top x < 0, w'^\top x \geq 0\}\|(w'^\top x - \sigma(v^\top x))x\|\Big]\end{aligned}$$

Note that we can bound the above three terms similarly as Vardi et al. (2021) by conditioning on the event in which $\|x\|$ is given, where $f_{\|x\|}$ is the p.d.f. of $\|x\|$.

$$
\begin{aligned}
&= \int \mathbb{E}\left[\|((w-w')^\top x)x\| \mid \|x\|\right] f_{\|x\|}dx \\
&+ \int \mathbb{E}\left[\mathbb{1}\{w^\top x \geq 0, w'^\top x < 0\}\|(w^\top x - \sigma(v^\top x))x\| \mid \|x\|\right] f_{\|x\|}dx \\
&+ \int \mathbb{E}\left[\mathbb{1}\{w^\top x < 0, w'^\top x \geq 0\}\|(w'^\top x - \sigma(v^\top x))x\| \mid \|x\|\right] f_{\|x\|}dx \\
&\leq \|w-w'\| \int \|x\|^2 f_{\|x\|}dx \\
&+ (\|w\|+\|v\|) \int \mathbb{E}\left[\mathbb{1}\{w^\top x \geq 0, w'^\top x < 0\} \mid \|x\|\right] \|x\|^2 f_{\|x\|}dx \\
&+ (\|w'\|+\|v\|) \int \mathbb{E}\left[\mathbb{1}\{w^\top x < 0, w'^\top x \geq 0\} \mid \|x\|\right] \|x\|^2 f_{\|x\|}dx
\end{aligned}
$$

We can directly bound both $\mathbb{P}\{w^\top x \geq 0, w'^\top x < 0 \mid \|x\|\}$ and $\mathbb{P}\{w^\top x < 0, w'^\top x \geq 0 \mid \|x\|\}$ by $4\beta_3/C_\ell \cdot \|w-w'\| \cdot \|x\|$ using the same argument in the proof of Lemma D.4 of Vardi et al. (2021), hence the above can be bounded as

$$
\begin{aligned}
&\leq \|w-w'\| \int \|x\|^2 f_{\|x\|}dx + 2(C_u + \sqrt{C_1^2+C_2^2}) \cdot \frac{4\beta_3}{C_\ell}\|w-w'\| \int \|x\|^3 f_{\|x\|}dx \\
&\leq (c_1' + \frac{8\beta_3(C_u+\sqrt{C_1^2+C_2^2})c_2'}{C_\ell}) \cdot \|w-w'\|
\end{aligned}
$$

for absolute constants c_1', c_2' as the second and third moments of $\|x\|$ due to properties of $O(1)$-regular distributions.

□

Lemma A.2 (Lemma D.5 in Vardi et al. (2021)). *Let $f : \mathbb{R}^{d+1} \to \mathbb{R}$ and $\ell > 0$. Assume for any $w, w' \in \mathbb{R}^{d+1}$ such that $\forall\ \lambda \in [0,1]$*

$$\|\nabla f((1-\lambda)w + \lambda w') - \nabla f(w)\| \leq \ell\lambda\|w'-w\|$$

then the following holds:

$$f(w') \leq f(w) + \langle \nabla f(w), w'-w\rangle + \frac{\ell}{2}\|w'-w\|^2$$

With all the above lemmas in place, we are now ready to prove Theorem 1.1.

Proof of Theorem 1.1 As described in Section 4.1, we inductively maintain two invariants in every iteration of the algorithm:

$$\text{(A)}\quad \|w_t - v\|_2 \leq O(1), \quad \text{and} \quad \text{(B)}\quad F(0) - F(w_t) = \Omega(1).$$

These two invariants are true at $t=0$ due to our initialization w_0. Lemma B.3 guarantees with at least constant probability $\Omega(1)$, both the invariants hold for w_0. The proof that both the invariants continue to hold follows from the progress made by the algorithm due to a decrease in both $\|w_t - v\|_2$ and $F(w_t)$ (note that we only need to show they do not increase to maintain the invariant).

The proof consists of three parts. For the first part, at time t, assuming $F(w_t) \leq F(0) - \delta$ holds, then by directly applying Lemma 4.5, we conclude that as long as $\|w_t - v\|^2 > C_p^2\gamma^{-1}(OPT + \varepsilon)$ for some constant $C_p > 0$, with learning rate $\eta = c_\eta d^{-1}$ where $c_\eta > 0$ is a constant, gradient descent always makes progress towards v.

In addition, since whenever $\|w_t - v\|^2 > C_p^2 \gamma^{-1}(OPT + \varepsilon)$, $\|w_t - v\|^2 - \|w_{t+1} - v\|^2$ is lower bounded by $\eta C'(OPT+\varepsilon)$ for some constant $C' > 0$, after $T = \|w_0 - v\|^2 C'^{-1} \eta^{-1}(OPT+\varepsilon)^{-1} \leq O(d(OPT + \varepsilon)^{-1})$ iterations we get $\|w_T - v\|^2 \leq O(OPT) + \varepsilon$, and by Lemma 4.4 this implies $F(w_T) \leq O(OPT) + \varepsilon$, therefore $L(w_T) \leq O(OPT) + \varepsilon$.

In the second part of the proof, we show that if w_0 is initialized such that $F(w_0) \leq F(0) - \delta$ for some $\delta > 0$, then while gradient descent is still iterating, the inequality $F(w_t) \leq F(w_0) \leq F(0) - \delta$ always holds.

By Lemma A.1 which establishes the smoothness of $\nabla F(w)$ between two iterates w and w', we can apply Lemma A.2 as

$$F(w') \leq F(w) + \langle \nabla F(w), w' - w \rangle + \frac{\ell}{2}\|w' - w\|^2$$

where $\ell = (c_1' + \frac{8\beta_3(C_u + \sqrt{C_1^2 + C_2^2})c_2'}{C_\ell})$. Note that the conditions in Lemma A.1 are met since at every timestep t, for some constant $C_\delta > 0$ $\|w_t\| \geq \frac{\sqrt{\delta}}{\sqrt{C_\delta \|v\|}} = C_\ell > 0$ implied by Lemma 4.1, and $\|w_t\| \leq \sqrt{C_1^2 + C_2^2} = C_u$ as well by assumption.

Now, substitute w with w_t and w' with $w_t - \eta \nabla L(w)$ yields

$$F(w_t - \eta \nabla L(w_t)) \leq F(w_t) - \eta \langle \nabla F(w_t), \nabla L(w_t) \rangle + \frac{\ell \eta^2}{2}\|\nabla L(w_t)\|^2$$

Note that

$$\langle \nabla F(w_t), \nabla L(w_t) \rangle = \langle \nabla F(w_t), \nabla F(w_t) + H(w_t) \rangle = \|\nabla F(w_t)\|^2 + \langle \nabla F(w_t), H(w_t) \rangle$$

where $H(w_t) = \mathbb{E}[(\sigma(v^\top x) - y)\sigma'(w_t^\top x)x]$.

Next, we define $u = \frac{\nabla F}{\|\nabla F\|}$. Note that $u \in \mathbb{R}^{d+1}$ is a fixed unit vector (it already involves an expectation over x); hence

$$\begin{aligned}
|\langle \nabla F(w_t), H(w_t) \rangle| &= \|\nabla F(w_t)\| \cdot \left| \left\langle \mathbb{E}\left[(\sigma(v^\top x) - y)\sigma'(w^\top x)x\right], u \right\rangle \right| \\
&= \|\nabla F(w_t)\| \left| \mathbb{E}\left[(\sigma(v^\top x) - y)\sigma'(w^\top x)u^\top x\right] \right| \leq \|\nabla F(w_t)\| \cdot \left| \mathbb{E}\left[(\sigma(v^\top x) - y)u^\top x\right] \right| \\
&\leq \|\nabla F(w_t)\| \, \mathbb{E}\left[\left|\sigma(v^\top x) - y\right| \cdot \left|u^\top x\right| \right] \leq \|\nabla F(w_t)\| \cdot \sqrt{OPT} \cdot \sqrt{\mathbb{E}\left[(u^\top x)^2\right]}
\end{aligned}$$

Note that

$$\mathbb{E}\left[(u^\top x)^2\right] = \mathbb{E}\left[(\widetilde{u}^\top \widetilde{x} + b_u)^2\right] \leq 2\,\mathbb{E}\left[(\widetilde{u}^\top \widetilde{x})^2\right] + 2b_u^2 = 2\Big(\beta_2 \|\widetilde{u}\|^2 + b_u^2\Big) \leq 2(\beta_2 + 1)$$

Therefore,

$$\begin{aligned}
&|\langle \nabla F(w_t), H(w_t) \rangle| \leq \|\nabla F(w_t)\| \cdot \sqrt{OPT} \cdot \sqrt{\mathbb{E}\left[(u^\top x)^2\right]} \leq \|\nabla F(w_t)\| \cdot \sqrt{2(\beta_2 + 1)OPT} \\
&\implies \langle \nabla F(w_t), H(w_t) \rangle \geq -\|\nabla F(w_t)\| \cdot \sqrt{2(\beta_2 + 1)OPT}
\end{aligned}$$

Plugging this back to the expression for $\langle \nabla F(w_t), \nabla L(w_t) \rangle$ yields

$$\begin{aligned}
\langle \nabla F(w_t), \nabla L(w_t) \rangle &= \|\nabla F(w_t)\|^2 + \langle \nabla F(w_t), H(w_t) \rangle \\
&\geq \|\nabla F(w_t)\|^2 - \|\nabla F(w_t)\| \cdot \sqrt{2(\beta_2 + 1)OPT} \\
&= \|\nabla F(w_t)\| \Big(\|\nabla F(w_t)\| - \sqrt{2(\beta_2 + 1)OPT} \Big) \\
&\geq \|\nabla F(w_t)\| \Big(\|\nabla F(w_t)\| - \sqrt{2(\beta_2 + 1)(OPT + \varepsilon)} \Big)
\end{aligned}$$

Since we have assumed that gradient descent is still in progress, implying $\|w_t - v\|$ is not at most $\sqrt{OPT+\varepsilon}$ yet, hence by Lemma 4.3 $\|\nabla F(w)\| > C_G\sqrt{OPT+\varepsilon}$ at this point, therefore

$$\langle \nabla F(w_t), \nabla L(w_t)\rangle \geq \|\nabla F(w_t)\|\Big(\|\nabla F(w_t)\| - \sqrt{2(\beta_2+1)(OPT+\varepsilon)}\Big) \geq 80(\beta_2+1)(OPT+\varepsilon)$$

with $C_G = 100(\beta_2+1)$; and by setting η and C_p as specified above, we have

$$-\eta\langle \nabla F(w_t), \nabla L(w_t)\rangle + \frac{\ell\eta^2}{2}\|\nabla L(w_t)\|^2 \leq -80\eta(\beta_2+1)(OPT+\varepsilon) + \frac{\ell\eta^2}{2}\|\nabla L(w)\|^2 \quad (21)$$

$$\leq \eta\Big(-80(\beta_2+1)(OPT+\varepsilon) + \frac{\ell\eta}{2}\cdot(4d\beta_2\|w_t - v\|^2 + 4dOPT)\Big) \quad (22)$$

$$= \eta\Big(-80(\beta_2+1)(OPT+\varepsilon) + 2d\beta_2\ell\eta\|w_t - v\|^2 + 2d\ell\eta(OPT+\varepsilon)\Big) \quad (23)$$

$$= \eta\Big((2d\ell\eta - 80(\beta_2+1))(OPT+\varepsilon) + 2d\beta_2\ell\eta\|w_t - v\|^2\Big) \quad (24)$$

$$\leq \eta\cdot\|w_t - v\|^2\cdot\Big(\frac{\gamma(2d\ell\eta - 80(\beta_2+1))}{C_p^2} + 2d\beta_2\ell\eta\Big) \leq 0 \quad (25)$$

by mandating $\beta_2 \geq \frac{\sqrt{3200\gamma\ell + (\ell C_p^2 - 800)^2} + \ell C_p^2}{1600}$. Note that this does not lose generality since we can always choose a suitable upper-bound for the second moment along any direction of the $O(1)$-regular distribution $\widetilde{\mathcal{D}}_x$. Hence, the above inequality implies

$$F(w_t - \eta\nabla L(w_t)) \leq F(w_t) \leq \ldots \leq F(w_0) \leq F(0) - \delta$$

Finally, in the last part of the proof, a direct application of Lemma B.3 justifies the assumption that w_0 can be initialized such that $F(w_0)$ is less than $F(0)$ by a constant amount with constant probability depending only on b_v; and since $|b_v| = O(1)$ by assumption, for absolute constants $c_1, c_2 > 0$, with probability at least c_2, $F(w_0) \leq F(0) - c_1^2$, which concludes the proof.

B Generalizing beyond Gaussian marginals

The above algorithmic result can be generalized to a broader class of marginals than Gaussians, that we call $O(1)$-regular marginals.

$O(1)$-regular marginals: Assumptions about the marginals over $\widetilde{x}$ We make the following assumptions about the marginal distribution $\widetilde{\mathcal{D}}_x$ over $\widetilde{x} \in \mathbb{R}^d$: there exists absolute constants $\beta_1, \beta_2', \beta_2, \beta_3, \beta_4, \beta_5 > 0$ and $\beta_0 : \mathbb{R}_+ \to \mathbb{R}_+$, such that

(i) Approximate isotropicity and bounded fourth moments: for every unit vector $u \in \mathbb{R}^d$, $\mathbb{E}_{\widetilde{x}\sim\widetilde{\mathcal{D}}_x}[\langle u, \widetilde{x}\rangle^2] \in [1/\beta_2', \beta_2]$, and $\mathbb{E}_{\widetilde{x}\sim\widetilde{\mathcal{D}}_x}[\langle u, \widetilde{x}\rangle^4] \leq \beta_4$.

(ii) Anti-concentration: there exists an absolute constant $\beta_3 > 0$ such that for every unit vector $\tilde{u} \in \mathbb{R}^d$ and $\delta > 0$,

$$\sup_{t\in\mathbb{R}} \mathbb{P}_{\widetilde{x}\sim\widetilde{\mathcal{D}}_x}\Big[\langle \tilde{u}, \widetilde{x}\rangle \in (t-\delta, t+\delta)\Big] \leq \min\{\beta_3\delta, 1\}.$$

(iii) Spread out: there exists $\beta_0 : \mathbb{R}_+ \to \mathbb{R}_+$ such that $\beta_0(|b_v|) > 0$ is a constant when $|b_v|$ is a constant, and

$$\forall \tilde{v} \in \mathbf{S}^{d-1}, \quad \mathop{\mathbb{E}}_{\widetilde{x}\sim\widetilde{\mathcal{D}}_x}\Big[\sigma(\tilde{v}^\top\widetilde{x} + b_v)\Big] \geq \beta_0(|b_v|).$$

(iv) 2-D projections: In every 2-dimensional subspace of $\mathbb{R}^d$ spanned by orthonormal unit vectors $\tilde{u}_1, \tilde{u}_2 \in \mathbb{R}^d$, we have a set $G_{\tilde{u}_1,\tilde{u}_2} \subset \mathbb{R}$ such that ,

$$\mathop{\mathbb{P}}_{\widetilde{x}\sim\widetilde{\mathcal{D}}_x}[\tilde{u}_2^\top\widetilde{x} \in G_{\tilde{u}_1,\tilde{u}_2}] = 1 - o(1), \text{ and} \quad (26)$$

$$\forall t \in G_{\tilde{u}_1,\tilde{u}_2}, \quad \mathop{\mathbb{E}}_{\widetilde{x}\sim\widetilde{\mathcal{D}}_x}\Big[\sigma(\tilde{u}_1^\top\widetilde{x}) \mid \tilde{u}_2^\top\widetilde{x} = t\Big] \geq \beta_5 \cdot \mathop{\mathbb{E}}_{\widetilde{x}\sim\widetilde{\mathcal{D}}_x}\big[\sigma(\tilde{u}_1^\top\widetilde{x})\big]. \quad (27)$$

In other words, the conditional expectation of $\sigma(\tilde{u}_1^\top \widetilde{x})$ is not much smaller after conditioning on the projection in an orthogonal direction $\tilde{u}_2$, for most values of $\tilde{u}_2^\top \widetilde{x}$. Note that for a Gaussian $N(0, I)$, the r.v.s $\tilde{u}_1^\top \widetilde{x}, \tilde{u}_2^\top \widetilde{x}$ are independent, so this condition holds with $\beta_5 = 1$ and $G_{\tilde{u}_1, \tilde{u}_2} = \mathbb{R}$.

We remark that Gaussian distribution $\mathcal{N}(0, I)$ is $O(1)$-regular i.e., all the constants $\beta_1, \beta_2, \beta_2', \beta_5 = 1, \beta_3 \leq 2$, and $\beta_0(b_v) = \mathbb{E}_{g \sim N(0,1)}[\sigma(g + b_v)] > 0$ for all $b_v \in (-\infty, \infty)$; in fact β_0 is an increasing function that is 0 only at $-\infty$. We also note that assumptions of this flavor have also been used in prior works including Vardi et al. (2021), which inspired parts of our analysis. In particular, Vardi et al. (2021) assume a lower-bound on the density for any 2-dimensional marginal; our assumption (4) on the 2-dimensional marginals is qualitatively weaker (it is potentially satisfied by even discrete distributions), and moreover we only need the condition to be satisfied for a large fraction of values of $\tilde{u}_2^\top \widetilde{x}$ (and not all). See Section B for the generalized version of our main theorem.

In this section, we present the main theorem as follows.

Theorem B.1 (Generalization beyond Gaussian marginals)**.** *For any absolute constants $C_1 \geq 1, C_2 > 0$, there exists absolute constants $c_3 > 0, c_\eta > 0$ such that the following holds. Let $\widetilde{\mathcal{D}}_x$ be a distribution over $(\widetilde{x}, y) \in \mathbb{R}^d \times \mathbb{R}$ where the marginal over $\widetilde{x}$ are regular with constant parameters $\beta_1, \beta_2', \beta_2, \beta_3, \beta_4, \beta_5$, and $\beta_0(b_v)$ as defined above. Let $H = \{(\widetilde{w}, b_w) : \|\widetilde{w}\| \in [1/C_1, C_1], |b_w| \leq C_2\}$, and consider population gradient descent iterates: $w_{t+1} = w_t - \eta \nabla L(w)$. For any $\varepsilon > 0$ and learning rate $\eta = c_\eta d^{-1}$, when starting from $w_0 = (\widetilde{w}_0, 0)$ where $\widetilde{w}_0$ is randomly initialized from a radially symmetric distribution, with at least constant probability $c_3 > 0$ one of the iterates w_T of gradient descent after poly$(d, 1/\varepsilon)$ steps satisfies $L(w_T) = O(OPT) + \varepsilon$.*

We now describe the generalization to regular distributions of the necessary lemmas for analyzing gradient descent in Section B.1.

B.1 Generalized Lemmas for Regular distributions

In the following lemma, similar to Lemma 4.5, we argue that throughout gradient descent, $\|w_t - v\|^2$ continues to decrease as long as $\|w_t - v\|$ is not too small.

Lemma B.2 (Decrease in $\|w_t - v\|$)**.** *Let $\widetilde{\mathcal{D}}_x$ be $O(1)$-regular with parameters defined above. Assume at time t, $F(w_t) \leq F(0) - \delta$ where $\delta > 0$ is a constant. For constants $C_p, C' > 0$ and γ defined as in Lemma 4.2, if for some $\varepsilon > 0$ $\|w_t - v\|^2 > \gamma^{-1} C_p^2 (OPT + \varepsilon)$, then $\|w_{t+1} - v\|^2 \leq \|w_t - v\|^2 - \eta C'(OPT + \varepsilon)$.*

Proof. Resembling the proof of Lemma 4.5, to lower-bound $\|w_t - v\|^2 - \|w_{t+1} - v\|^2$, we will give a lower bound for $\langle \nabla L(w_t), w_t - v \rangle$ and an upper bound for $\|\nabla L(w_t)\|^2$.

Lower bounding $\langle \nabla L(w_t), w_t - v \rangle$ A direct application of Lemma 4.2 already gives a lower bound on $\langle \nabla F(w_t), w_t - v \rangle$, hence we need only focus on lower bounding $\langle \mathbb{E}[(\sigma(v^\top x) - y)\sigma'(w_t^\top x)x], w_t - v \rangle$, and by Cauchy–Schwarz and Young's inequality, we immediately get

$$\begin{aligned}
&\langle \mathbb{E}[(\sigma(v^\top x) - y)\sigma'(w_t^\top x)x], w_t - v \rangle \;=\; \mathbb{E}[(\sigma(v^\top x) - y)\sigma'(w^\top x)(w_t^\top x - v^\top x)] \\
&\geq -\sqrt{\mathbb{E}[(\sigma(v^\top x) - y)^2]}\sqrt{\mathbb{E}[(w_t^\top x - v^\top x)^2 \sigma'(w_t^\top x)]} \geq -\sqrt{2OPT}\sqrt{\mathbb{E}[((w_t - v)^\top x)^2]} \\
&\geq -\sqrt{2OPT} \cdot \sqrt{C_\beta \beta_2}\|w_t - v\| \geq -C_\beta' \sqrt{\beta_2} \cdot \sqrt{OPT} \cdot \|w_t - v\|
\end{aligned}$$

with constants $C_\beta, C_\beta' > 0$. Putting the bound above along with that of Lemma 4.2 together we get

$$\nabla L(w_t) = \nabla F(w_t) + \mathbb{E}[(\sigma(v^\top x) - y)\sigma'(w_t^\top x)x] \geq \gamma \|w_t - v\|_2^2 - C_\beta' \sqrt{\beta_2} \cdot \sqrt{OPT} \cdot \|w_t - v\|_2$$

Upper bounding $\|\nabla L(w_t)\|^2$ Recall $\nabla L(w_t) = \nabla F(w_t) + H(w_t) \implies \|\nabla L(w_t)\|^2 \leq 2\|\nabla F(w_t)\|^2 + 2\|H(w_t)\|^2$. For the first term,

$$\|\nabla F(w_t)\| \leq \; \mathbb{E}[|\sigma(w_t^\top x) - \sigma(v^\top x)| \cdot |\sigma'(w_t^\top x)| \cdot \|x\|] \leq \; \mathbb{E}[|w_t^\top x - v^\top x| \cdot \|x\|]$$

since $\sigma(\cdot)$ is 1-Lipschitz (i.e. $|\sigma(z) - \sigma(z')| \le |z - z'|$) and $\sigma'(\cdot) \le 1$. Hence, applying Cauchy-Schwarz yields

$$\le \sqrt{\mathbb{E}[|w_t^\top x - v^\top x|^2] \cdot \mathbb{E}[\|x\|^2]} \le \|w_t - v\| \cdot \sqrt{\beta_2 d + 1}$$

Similarly, for the second term,

$$\|H(w_t)\| \le \mathbb{E}[|\sigma(v^\top x) - y| \cdot \|x\|] \le \sqrt{\mathbb{E}[|\sigma(v^\top x) - y|^2] \cdot \mathbb{E}[\|x\|^2]} \le \sqrt{2OPT} \cdot \sqrt{\beta_2 d + 1}$$

Using the above two expression, we can hence bound $\|\nabla L(w_t)\|^2$ as

$$\|\nabla L(w_t)\|^2 \le 2\|F(w_t)\|^2 + 2\|H(w_t)\|^2 \le C''_\beta d\|w - v\|^2 + C''_\beta dOPT$$

for some constant $C''_\beta > 0$.

Lower bounding $\|w_t - v\|^2 - \|w_{t+1} - v\|^2$ The above inequalities yield

$$\begin{aligned}
&\|w_t - v\|^2 - \|w_{t+1} - v\|^2 = 2\eta\langle \nabla L(w_t), w_t - v\rangle - \eta^2\|\nabla L(w_t)\|^2 \\
&\ge 2\eta \cdot \Big[\gamma\|w_t - v\|^2 - C'_\beta\sqrt{\beta_2}\sqrt{OPT}\|w_t - v\|\Big] - C''_\beta d\eta^2 \cdot (\|w - v\|^2 + OPT) \\
&\ge 2\eta \cdot \Big[\gamma\|w_t - v\|^2 - C'_\beta\sqrt{\beta_2}\gamma^{1/2}C_p^{-1}\|w_t - v\|^2\Big] - C''_\beta d\eta^2 \cdot (\|w - v\|^2 + OPT) \\
&= 2\eta\Big(\gamma - C'_\beta\sqrt{\beta_2}\gamma^{1/2}C_p^{-1} - \frac{C''_\beta}{2}d\eta\Big)\|w_t - v\|^2 - 2\eta \cdot \frac{C''_\beta}{2}d\eta OPT
\end{aligned}$$

due to our assumption that (b) does not hold yet, i.e. $\|w_t - v\| > C_p\gamma^{-1/2}\sqrt{(OPT + \varepsilon)} > C_p\gamma^{-1/2}\sqrt{OPT}$ with some constant $C_p > 0$, implying $\sqrt{OPT} < \gamma^{1/2}C_p^{-1}\|w_t - v\|$. Consequently, by choosing $\eta \le O(d^{-1})$, we get

$$\begin{aligned}
&\ge 2\eta\Big(C_1\gamma\|w_t - v\|^2 - C_2 OPT\Big) \ge 2\eta\Big(C_1C_p^2(OPT + \varepsilon) - C_2 OPT\Big) \\
&\ge \eta C'(OPT + \varepsilon)
\end{aligned}$$

where $C_1, C_2, C' > 0$ are constants. Hence the proof follows. □

With the lemmas above, we are now ready to prove Theorem B.1.

B.2 Proof of Theorem B.1

The proof highly resembles that of Theorem 1.1 by inductively maintaining the same two invariants in every iteration of the algorithm:

$$\text{(A)} \quad \|w_t - v\|_2 \le O(1), \quad \text{and} \quad \text{(B)} \quad F(0) - F(w_t) = \Omega(1).$$

Hence, we only highlight the difference compared to the previous proof.

The proof also consists of three parts. For the first part, we simply replace Lemmas 4.5 and 4.4 with Lemmas B.2 and 4.4, resulting in the same argument that after $T \le O(d(OPT + \varepsilon)^{-1})$ iterations we get $F(w_T) \le O(OPT) + \varepsilon$, therefore $L(w_T) \le O(OPT) + \varepsilon$.

In the second part of the proof, Lemmas 4.3, A.1, A.2 remain valid for $O(1)$-regular distributions, therefore we need only note that for any unit vector $u \in \mathbb{R}^{d+1}$,

$$\mathbb{E}[(u^\top x)^2] \le 2\,\mathbb{E}[(\widetilde{u}^\top \widetilde{x})^2] + 2b_u^2 \le 2\beta_2 + 2b_u^2 \le O(\beta_2)$$

which only affects the bounds for $\|\nabla L(w_t)\|$ and $\|H(w_t)\|$ up to a constant factor. Hence the inequality $F(w_t - \eta\nabla L(w_t)) \le F(w_t) \le F(w_0) \le F(0) - \delta$ also holds.

Finally, in the last part of the proof, a direct application of Lemma B.3 justifies the initialization assumption, which concludes the proof.

B.3 RANDOM INITIALIZATION

We now prove the initialization lemma assuming weak conditions on the marginal distribution over $\widetilde{x} \in \mathbb{R}^d$ which is $\widetilde{\mathcal{D}}_x$ (recall that the standard Gaussian $N(0, I)$ also satisfies all of the properties). We will initialize $w = (\tilde{w}, b_w)$ with $b_w = 0$ and $\tilde{w}$ drawn from a spherical symmetric distribution $\mathcal{D}_w$. $\mathcal{D}_w$ first picks the length $\rho \in \mathcal{D}_\rho$, and then sets $\tilde{w} = \rho\hat{w}$, where $\hat{w}$ is a uniformly random unit vector. The distribution $\mathcal{D}_\rho$ can be any distribution that is reasonably spread out – it just needs to place non-negligible probability in any constant length interval $(a_1\|\tilde{v}\|_2, a_2\|\tilde{v}\|_2)$ where $a_2 > a_1 > 0$ are constants.

As stated in the preliminaries, we assume for simplicity that $\|\tilde{v}\|_2 = 1$ (or $\Theta(1)$); this is essentially the same as assuming that we know the length scale of $\|\tilde{v}\|_2$, since we can scale the input by this length $\|\tilde{v}\|_2$ (see Proposition C.1). Please refer to Lemma 5.1 when we do not know the length scale of $\|\tilde{v}\|_2$. For convenience, we will set $\mathcal{D}_\rho$ to be the absolute value of a standard Gaussian $N(0, 1)$ (or $N(0, \beta^2)$ with $\beta \in [1, 2]$.

Lemma B.3. *There exists $c_1(v), c_2(v), c_3(v) > 0$ which only depend on $b_v/\|\tilde{v}\|_2$ (and not on the dimension), and are both absolute constants when $|b_v|/\|\tilde{v}\|_2 = O(1)$, such that the following holds. When $w = (\tilde{w}, b_w = 0)$ is drawn according to $\tilde{w} = \rho\|\tilde{v}\|_2\hat{w} \sim \mathcal{D}_w$ described above (with $\hat{w}$ being a uniformly random unit vector, and $\rho \sim \mathcal{D}_\rho$ being the absolute value of a normal $N(0, \beta^2)$ with $\beta \in [1, 2]$). Then with probability at least $c_2(v) > 0$, we have*

$$F(w) \le F(0) - c_1(v)^2\|\tilde{v}\|_2^2, \text{ and } \|w - v\| \le c_3(v)\|\tilde{v}\|_2. \tag{28}$$

In the above lemma, if $\widetilde{\mathcal{D}}_x$ is a standard Gaussian $N(0, I)$, it suffices to choose for example $c_1(v) = c_0\,\mathbb{E}_{g_1\sim N(0,1)}[\sigma(g_1 + b_v/\|\tilde{v}\|)]\Big) = c_0 \cdot \left(\frac{b_v}{\|\tilde{v}\|}\Phi(\frac{b_v}{\|\tilde{v}\|}) + \frac{1}{\sqrt{2\pi}}e^{-b_v^2/2\|\tilde{v}\|^2}\right)$ for some universal constants $c_0, c_0', c_0'' > 0$. $c_2(v)$ and $c_3(v)$ are also chosen similarly as constants that only depend on $|b_v|/\|\tilde{v}\|$ and not on any dimension dependent term. We remark that for random initialization to work, we only need the probability of success $\eta > 0$ to be non-negligible (e.g., at least an inverse polynomial). We can always try $O(1/\eta)$ many random initializers, and amplify the success probability to be at least 0.99.

Proof. For convenience we define $\widehat{b}_v := b_v/\|\tilde{v}\|_2, \widehat{v} := v/\|\tilde{v}\|_2$, so they are normalized w.r.t. the length of $\tilde{v}$. The conditions of the lemma assume that $|\widehat{b}_v| = O(1)$.

By definition, the distribution of $\tilde{w} \in \mathbb{R}^d$ is spherically symmetric.

$$\begin{aligned}
F(w) - F(0) &= \frac{1}{2}\mathop{\mathbb{E}}_x\left[(\sigma(\tilde{w}^\top x) - \sigma(\tilde{v}^\top x + b_v))^2\right] - \frac{1}{2}\mathop{\mathbb{E}}_x\left[\sigma(\tilde{v}^\top x + b_v))^2\right] \\
&= \frac{1}{2}\mathop{\mathbb{E}}_x\left[(\sigma(\tilde{w}^\top x)^2\right] - \mathop{\mathbb{E}}_x\left[\sigma(\tilde{w}^\top x)\sigma(\tilde{v}^\top x + b_v))\right] \\
&= \frac{\rho^2\|\tilde{v}\|_2^2}{2}\mathop{\mathbb{E}}_x\left[(\sigma(\widehat{w}^\top x)^2\right] - \rho\|\tilde{v}\|_2^2\mathop{\mathbb{E}}_x\left[\sigma(\widehat{w}^\top x)\sigma(\widehat{v}^\top x + \widehat{b}_v))\right],
\end{aligned}$$

where $\tilde{w} = \rho\|\tilde{v}\|_2\widehat{w}$ with $\widehat{w}$ being the unit vector along $\tilde{w}$. For a fixed $\rho \in \mathbb{R}_+$, $\widehat{w}$ (and hence $\tilde{w}$) is picked along a uniformly random direction i.e., $\widehat{w} \sim_U \mathbb{S}^{d-1}$. Hence for $x \sim \widetilde{\mathcal{D}}_x$

$$\begin{aligned}
\mathop{\mathbb{E}}_{\widehat{w}\sim\mathbb{S}^{d-1}}[F((\rho\widehat{w}, 0)) - F(0)] &= \frac{\rho^2\|\tilde{v}\|_2^2}{2}\mathop{\mathbb{E}}_{\widehat{w}\sim_U\mathbb{S}^{d-1}}\mathop{\mathbb{E}}_{x\sim\widetilde{\mathcal{D}}_x}\left[(\sigma(\widehat{w}^\top x)^2\right] \qquad (29) \\
&\quad - \rho\|\tilde{v}\|_2^2\mathop{\mathbb{E}}_{\widehat{w}\sim_U\mathbb{S}^{d-1}}\mathop{\mathbb{E}}_{x\sim\widetilde{\mathcal{D}}_x}\left[\sigma(\widehat{w}^\top x)\sigma(\widehat{v}^\top x + \widehat{b}_v))\right] \\
&= \|\tilde{v}\|_2^2\big(c'\rho^2 - 2c_3(v)\rho\big), \qquad (30)
\end{aligned}$$

where $c' > 0$ is a universal constant based on our assumptions about $\widetilde{\mathcal{D}}_x$ ($c' = 0.5$ for $x \sim N(0, I)$).

We now derive an expression for $c_3(v)$, and prove that it is a constant independent of the dimension. Let $\widehat{w} = z_1\widehat{v} + z_2 w^\perp$ where $w^\perp$ is some unit vector orthogonal to $\tilde{v}$. Note that z_1, z_2 are r.v.s that depend only on the choice of the initializer (our rotationally invariant distribution), and not on $\widetilde{\mathcal{D}}_x$. For $\widehat{w} \sim_U \mathbb{S}^{d-1}$, the typical values $\mathbb{E}[z_1^2] = 1/d$ and $\mathbb{E}[z_2^2] = 1 - 1/d$; moreover z_1 and z_2 are

symmetric (around 0), and their signs are independent. Let the r.v.s $\xi_1 = \langle \widetilde{x}, \widehat{v} \rangle, \xi_2 = \langle \widetilde{x}, w^\perp \rangle$ denote the marginal distribution along $\widehat{v}, w^\perp$. The ξ_1, ξ_2 are independent of z_1, z_2 (but ξ_1, ξ_2 could be dependent); these also satisfy condition (3) about the 2-dimensional marginals of $\widetilde{\mathcal{D}}_x$ because it is $O(1)$-regular.

$$\begin{aligned} c_3(v) &= \mathop{\mathbb{E}}_{\widehat{w} \sim_U \mathbb{S}^{d-1}} \mathop{\mathbb{E}}_{x \sim \widetilde{\mathcal{D}}_x} \Big[\sigma(\widehat{w}^\top x) \sigma(\widehat{v}^\top x + \widehat{b}_v)) \Big] \\ &= \mathop{\mathbb{E}}_{z_1, z_2} \mathop{\mathbb{E}}_{\xi_1, \xi_2} \Big[\sigma(z_1 \xi_1 + z_2 \xi_2) \sigma(\xi_1 + \widehat{b}_v) \Big] \\ &= \mathop{\mathbb{E}}_{z_1, z_2} \mathop{\mathbb{E}}_{\xi_1, \xi_2} \Big[\sigma(z_1 \xi_1 + z_2 \xi_2) \sigma(\xi_1 + \widehat{b}_v) \Big]. \end{aligned}$$

Since z_1 is a symmetric r.v.,

$$\begin{aligned} c_3(v) &= \frac{1}{2} \mathop{\mathbb{E}}_{z_1, z_2} \mathop{\mathbb{E}}_{\xi_1, \xi_2} \Big[\sigma(|z_1 \xi_1| + z_2 \xi_2) \sigma(\xi_1 + \widehat{b}_v) \Big] + \frac{1}{2} \mathop{\mathbb{E}}_{z_1, z_2} \mathop{\mathbb{E}}_{\xi_1, \xi_2} \Big[\sigma(-|z_1 \xi_1| + z_2 \xi_2) \sigma(\xi_1 + \widehat{b}_v) \Big] \\ &\geq \frac{1}{2} \mathop{\mathbb{E}}_{z_1, z_2} \mathop{\mathbb{E}}_{\xi_1, \xi_2} \Big[\sigma(|z_1 \xi_1| + z_2 \xi_2) \sigma(\xi_1 + \widehat{b}_v) \Big] \\ &\geq \frac{1}{2} \mathop{\mathbb{E}}_{\xi_1, \xi_2} \mathop{\mathbb{E}}_{z_1, z_2} \Big[\sigma(z_2 \xi_2) \sigma(\xi_1 + \widehat{b}_v) \Big] \\ &\geq \frac{\mathbb{E}_z |z_2|}{2} \int_{t=-\widehat{b}_v}^{\infty} p(\xi_1 = t) \cdot (t + \widehat{b}_v) \cdot \mathop{\mathbb{E}}_{\xi_2} [\sigma(\xi_2) | \xi_1 = t] \, dt \quad \text{(since } z_2 \text{ is independent of } \xi_1, \xi_2) \\ &\geq \frac{1}{8} \int_{t=-\widehat{b}_v}^{\infty} p(\xi_1 = t) \cdot (t + \widehat{b}_v) \cdot \mathop{\mathbb{E}}_{\xi_2} [\sigma(\xi_2) | \xi_1 = t] \, dt \quad \text{(since } z_2 \text{ is independent of } \xi_1, \xi_2) \end{aligned}$$

since $\mathbb{E}[|z_2|] \geq 1/4$ (in fact when d is large, $|z_2| = 1 - O(1/\sqrt{d})$ for w.h.p.). We now split up the inner integral over $t \in [-\widehat{b}_v, \infty)$ into two parts depending on whether $\mathbb{E}_{\xi_2}[\sigma(\xi_2)|\xi_1 = t] \geq \beta_5 \, \mathbb{E}_{\xi_2}[\sigma(\xi_2)]$ is satisfied or not. Let Bad $\subset [-\widehat{b}_v, \infty)$ be subset where it is not satisfied. Note that from regularity of $\widetilde{\mathcal{D}}_x$, we have that $\mathbb{P}[\text{Bad}] = o(1)$. We only take the contribution from $t \in [-\widehat{b}_v, \infty) \setminus \text{Bad}$:

$$\begin{aligned} c_3(v) &\geq \frac{\beta_5}{8} \Big(\int_{t=-\widehat{b}_v}^{\infty} p(\xi_1 = t) \cdot (t + \widehat{b}_v) \cdot \mathop{\mathbb{E}}_{\xi_2} [\sigma(\xi_2)] \, dt - \int_{t=-\widehat{b}_v}^{\infty} p(\xi_1 = t) \mathbf{1}[t \in \text{Bad}] \cdot (t + \widehat{b}_v) \cdot \beta_5 \mathop{\mathbb{E}}_{\xi_2} [\sigma(\xi_2)] \, dt \Big) \\ &\geq \frac{\beta_5 \, \mathbb{E}_{\xi_2}[\sigma(\xi_2)]}{8} \Big(\mathop{\mathbb{E}}_{\xi_1} [\sigma(\xi_1 + \widehat{b}_v)] - \sqrt{\mathbb{P}[\text{Bad}] \cdot \int_{t \in \text{Bad}} p(\xi_1 = t) \cdot (t + \widehat{b}_v)^2 \, dt} \Big) \\ &\geq \frac{\beta_5 \beta_0(0)}{8} \Big(\mathop{\mathbb{E}}_{\xi_1} [\sigma(\xi_1 + \widehat{b}_v)] - \mathbb{P}[\text{Bad}]^{1/2} \cdot \Big(2 \int_{t \in \mathbb{R}} p(\xi_1 = t) \cdot (t^2 + \widehat{b}_v^2) \, dt \Big)^{1/2} \Big) \\ &\geq \frac{\beta_5 \beta_0(0)}{8} \Big(\beta_0(|\widehat{b}_v|) - o(1) \cdot \sqrt{2(\beta_2 + \widehat{b}_v^2)} \Big) \\ &\geq c_1 \beta_0(|\widehat{b}_v|), \end{aligned}$$

as required for an absolute constant $c_1 > 0$. Note that the last line used regularity to say $\beta_5 = \Omega(1)$ and lower bound $\mathbb{E}[\sigma(\widehat{v}^\top \widetilde{x} + \widehat{b}_v)] \geq \beta_0(|\widehat{b}_v|)$.

We now prove that the first part (31) holds with non-negligible probability. From (30), we note that for any $\rho \in \left[\frac{c_3(v)}{2c'}, \frac{c_3(v)}{c'} \right]$, we have that

$$\mathop{\mathbb{E}}_{\widehat{w} \sim_U \mathbb{S}^{d-1}} [F((\rho \widehat{w}, 0))] \leq F(0) - \|\tilde{v}\|_2^2 \frac{c_3(v)^2}{2c'}.$$

Moreover ρ is distributed as the absolute value of a standard normal with variance in $[1, 4]$; so we get that $\rho \in \left(\frac{c_3(v)}{2c'}, \frac{c_3(v)}{c'} \right)$ with probability at least $c_5(v) := \frac{1}{2\sqrt{2\pi}c'} \cdot e^{-\Omega(c_3(v)^2)} c_3(v)$, which is constant when $|\widehat{b}_v|$ is a constant.

Now we condition on the event that $\rho \in \left[\frac{c_3(v)}{2c'}, \frac{c_3(v)}{c'} \right]$. For a fixed ρ in this interval, let Z be a r.v. that captures the distribution of $F((\rho \|\tilde{v}\| \widehat{w}, 0)) - F(0)$ as $\widehat{w}$ is drawn uniformly from the unit sphere

$\mathbb{S}^{d-1}$. Note that $\mathbb{E}[Z] \le -\|\tilde{v}\|_2^2 c_3(v)^2/2c'$.

$$\begin{aligned}\mathrm{Var}[Z] &\le \mathbb{E}[F((\rho\|\tilde{v}\|_2\widehat{w},0))^2] \le 16 \mathop{\mathbb{E}}_{x\sim\widetilde{\mathcal{D}}_x}[\|\tilde{v}\|^4\sigma(\rho\widehat{w}^\top x)^4] + 16 \mathop{\mathbb{E}}_{x\sim\widetilde{\mathcal{D}}_x}[\|\tilde{v}\|^4\sigma(\widehat{v}^\top x + \widehat{b}_v)^4] \\ &\le 256\|\tilde{v}\|^4\Big(2\beta_4 + \widehat{b}_v^4\Big).\end{aligned}$$

Further for $\lambda = -\mathbb{E}[Z]/2$, we have from the Cantelli-Chebychev one-sided tail inequality we have for some absolute constant $c_6 > 0$

$$\mathbb{P}\Big[Z \le \mathbb{E}[Z]/2\Big] \ge \frac{\mathbb{E}[Z]^2}{4\,\mathrm{Var}[Z] + \mathbb{E}[Z]^2} \ge \min\Big\{c_6 c_3(v)^2/(\beta_4 + \widehat{b}_v^4), \frac{1}{2}\Big\} =: c_6(v),$$

where $c_6(v)$ is a constant when $\widehat{b}_v$ is a constant. This allows us to conclude that $F(w) < F(0) - \Omega(\|\tilde{v}\|^2)$ with probability at least $c_5(v)\cdot c_6(v)$ which is a constant when $\widehat{b}_v$ is a constant. Finally the $\|w - v\|_2 \le \|w\|_2 + \|\tilde{v}\|_2$ is upper bound just because of our choice of ρ and $\|\tilde{v}\|_2$ being upper bounded by assumption.

□

B.4 Multiscale Random Initialization (for Unknown Length Scale)

Lemma B.3 shows that if we guess the correct length scale of $\|\tilde{v}\|_2$ up to a factor of 2, then the random spherically symmetric initialization in Section 5 succeeds with constant probability. When we have unknown length scale $\|\tilde{v}\|_2 \in [1/M, M]$, the random initialization can try out the different length scales in geometric progression i.e., the length scale τ is chosen uniformly at random from $\{2^{-j} : j \in \mathbb{Z}, -\log M \le j \le \log M\}$.

Random initialization for unknown length scale We will initialize $w = (\tilde{w}, b_w)$ with $b_w = 0$ and $\tilde{w}$ drawn from a spherical symmetric distribution $\mathcal{D}_w$. The length is chosen from the distribution $\mathcal{D}_\rho$ so that it has a non-negligible probability in any constant length interval $(a_1\|v\|_2, a_2\|v\|_2)$ where $a_2 > a_1 > 0$ are constants: our specific choice picks the correct length scale with non-negligible probability, and is reasonably spread out.

We are given a parameter M such that $\|v\|_2 \in [2^{-\log M}, 2^{\log M}]$ (note that M can have large dependencies on d and other parameters; our guarantees will be polynomial in $\log M$). A random initializer $w = (\tilde{w}, 0)$ is drawn from $\mathcal{D}_{\text{unknown}}(M)$ as follows:

1. Pick j uniformly at random from $\big\{-\lceil\log M\rceil, -\lceil\log M\rceil + 1, \ldots, -1, 0, 1, \ldots, \lceil\log M\rceil\big\}$.
2. $\rho \in \mathbb{R}_+$ is drawn according to $\mathcal{D}_\rho$ as follows: we first pick[2] $g \sim N(0,1)$ and set $\rho = 2^j|g|$.
3. A uniformly random *unit* vector $\hat{w} \in \mathbb{R}^d$ is drawn and we output $\tilde{w} = \rho\hat{w}$. The initializer is $(\tilde{w}, 0)$.

We prove the following claim about the random initializer.

Lemma 5.1 *There exists $c_1(v), c_2(v), c_3(v) > 0$ which only depend on $b_v/\|\tilde{v}\|_2$ (and not on the dimension), and are both absolute constants when $|b_v|/\|\tilde{v}\|_2 = O(1)$, such that the following holds. When $w = (\tilde{w}, b_w = 0)$ is drawn according to the distribution $\mathcal{D}_{unknown}(M)$ described above for some given $M \ge 1$ satisfying $\|v\|_2 \in [1/M, M]$. Then with probability at least $c_2(v)/\log M$, we have*

$$F(w) \le F(0) - c_1(v)^2\|\tilde{v}\|_2^2, \text{ and } \|w - v\| \le c_3(v)\|\tilde{v}\|_2. \tag{31}$$

Proof. Since $\|\tilde{v}\|_2 \in [1/M, M]$, the random initialization will pick j^* with probability at least $1/(2\log M)$ such that $\|\tilde{v}\|_2 \in [2^{j^*}, 2^{j^*+1}]$. For this choice of j^*, we can apply Lemma B.3 (note that we only need a guess of $\|\tilde{v}\|_2$ up to a factor of 2) to get the required guarantee. □

[2] One can pick many other spread out distributions in place of the absolute value of a Gaussian.

C INVARIANCE TO SCALING

In this section we show that the guarantees of gradient descent do not change by scale the instance by a multiplicative factor of α. Here the instance is scaled by only multiplying the y values by the same factor α (but *not* scaling the point x). This allows us to assume that $\|\tilde{v}\|_2 = 1$ without loss of generality as long as the initializer is also in the same length scale (see Lemma 5.1 how the random initialization finds the correct length scale with reasonable probability).

Recall that we consider the loss function L which given hypothesis $w = (\tilde{w}, b_w) \in \mathbb{R}^{d+1}$ and input distribution $\tilde{\mathcal{D}}$ over $(\widetilde{x}, y) \in \mathbb{R}^d \times \mathbb{R}$ is

$$L(w, \tilde{\mathcal{D}}) = \tfrac{1}{2} \mathop{\mathbb{E}}_{(\widetilde{x},y)\sim\tilde{\mathcal{D}}}[(\sigma(\tilde{w}^\top \widetilde{x} + b_w) - y)^2]. \tag{32}$$

We show the following simple proposition.

Proposition C.1. *Let $\alpha > 0$, and let $\tilde{\mathcal{D}}$ be any distribution over $(\widetilde{x}, y) \in \mathbb{R}^d \times \mathbb{R}$, let $\tilde{\mathcal{D}}_\alpha$ be the corresponding distribution given by $(\widetilde{x}, y' = \alpha y)$ (only the y values are scaled). For every $w = (\tilde{w}, b_w) \in \mathbb{R}^{d+1}$ we have that*

$$L(\alpha w, \tilde{\mathcal{D}}_\alpha) = \alpha^2 \cdot L(w, \tilde{\mathcal{D}}), \quad \textit{where } \alpha w = (\alpha\tilde{w}, \alpha b_w). \tag{33}$$

Moreover, for two runs of gradient descent (with the same step size η) producing iterates $w_0, w_1, \ldots, w_T$ when run on $\tilde{\mathcal{D}}$ and producing iterates $w'_0, w'_1, \ldots, w'_T$ when run on $\tilde{\mathcal{D}}_\alpha$, we have:

$$\textit{if } w'_0 = \alpha w_0, \quad \textit{then} \quad \forall t \in \{0, 1, 2, \ldots, T\},\ w'_t = \alpha \cdot w_t. \tag{34}$$

Finally, if OPT and OPT_α are the optimal losses for $\tilde{\mathcal{D}}$ and $\tilde{\mathcal{D}}_\alpha$ respectively, then for any $\beta > 0$, $F(w_t) \le \beta \cdot OPT$ if and only if $F(w'_t) \le \beta \cdot OPT_\alpha$.

Proof. The first part follows directly from (32). We have

$$\begin{aligned} L(\alpha w, \tilde{\mathcal{D}}_\alpha) &= \tfrac{1}{2} \mathop{\mathbb{E}}_{(\widetilde{x},y')\sim\tilde{\mathcal{D}}_\alpha} [(\sigma(\alpha\tilde{w}^\top \widetilde{x} + \alpha b_w) - y')^2] = \tfrac{1}{2} \mathop{\mathbb{E}}_{(\widetilde{x},y)\sim\tilde{\mathcal{D}}} [(\sigma(\alpha\tilde{w}^\top \widetilde{x} + \alpha b_w) - \alpha y)^2] \\ &= \tfrac{1}{2} \mathop{\mathbb{E}}_{(\widetilde{x},y)\sim\tilde{\mathcal{D}}} [(\alpha\sigma(\tilde{w}^\top \widetilde{x} + b_w) - \alpha y)^2] = \alpha^2 L(w, \tilde{\mathcal{D}}). \end{aligned}$$

The second part uses the form of the gradient update through a simple induction. The base case is true since by assumption $w'_0 = \alpha w_0$. Suppose $w'_t = \alpha w_t$. Let $\mathcal{D}_\alpha$ denote the distribution over $x = (\widetilde{x}, 1), y' = \alpha y$ corresponding to $\tilde{\mathcal{D}}_\alpha$. Recall that $w'_{t+1} = w'_t - \nabla L(w'_t, \mathcal{D}_\alpha)$ where

$$\begin{aligned} \nabla L(w'_t, \mathcal{D}_\alpha) &= \mathop{\mathbb{E}}_{(x,y')\sim\mathcal{D}_\alpha} \Big[(\sigma(w^\top x) - y')\sigma'(w^\top x)x\Big]. \\ \text{Hence } w'_{t+1} &= w'_t - \eta\nabla L(w'_t, \mathcal{D}_\alpha) = w'_t - \eta \mathop{\mathbb{E}}_{(x,y')\sim\mathcal{D}_\alpha} \Big[(\sigma(\alpha w^\top x) - y')\sigma'(\alpha w^\top x)x\Big] \\ &= \alpha w_t - \alpha \cdot \eta \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \Big[(\sigma(\alpha w^\top x) - \alpha y)\sigma'(w^\top x)x\Big] \\ &= \alpha w_t - \alpha \cdot \nabla L(w_t, \mathcal{D}) = \alpha w_{t+1}. \end{aligned}$$

Note that the last but second line used the fact that $\sigma'(\alpha w^\top x) = \mathbf{I}[\alpha w^\top x \ge 0] = \sigma'(w^\top x)$ when $\alpha > 0$. The last part of the proposition just follows from the first claim that $L(\alpha w, \tilde{\mathcal{D}}_\alpha) = \alpha^2 \cdot L(w, \tilde{\mathcal{D}})$ for all w applied to $w_T, w'_T = \alpha w_T$ and the optimal solutions corresponding to OPT and OPT_α. □

Remark. We remark that the above proposition essentially shows that we can assume that $\|\tilde{v}\|_2 = 1$, almost without loss of generality. However, this proposition assumes that initializer $\tilde{w}_0$ can also be scaled accordingly i.e., the initializer $\tilde{w}_0$ continues to have the same length scale as $\tilde{v}$. This is achieved by our random initialization strategy in Section B.4, since it tries out many different length scales.

D Analysis of Gradient Descent with Finite Samples

In this section, we analyze gradient descent when trained on finite number of i.i.d. samples $(x_i, y_i) \sim D$. As in the previous sections, we assume the marginal distribution of x is a standard Gaussian. We will utilize the notations below, which are analogous to those defined with respect to the data distribution

- $\widehat{L}(w) = \frac{1}{2n}\sum_{i=1}^{n}(\sigma(w^\top x_i) - y_i)^2$
- $\widehat{F}(w) = \frac{1}{2n}\sum_{i=1}^{n}(\sigma(w^\top x_i) - \sigma(v^\top x_i))^2$
- $\widehat{H}(w) = \frac{1}{n}\sum_{i=1}^{n}(\sigma(v^\top x_i) - y_i)\sigma'(w^\top x_i)x_i$

where n is the number of samples.

Since we can only access n samples, we update the weight through full-batch gradient descent as follows

$$w_{t+1} = w_t - \eta\nabla\widehat{L}(w_t)$$

We are now ready to analyze gradient descent on finite samples. We first state the main result established in this section

Theorem D.1. *Let $C_1 \geq 1, C_2 > 0, c_3' > 0$ be absolute constants. Let D be a distribution over $(\widetilde{x}, y) \in \mathbb{R}^d \times \mathbb{R}$ where the marginal over $\widetilde{x}$ is the standard Gaussian $\mathcal{N}(0, I)$ and the distribution of y satisfies $|y| \leq B_Y$ for some $B_Y \geq 1$. Let $H = \{w = (\tilde{w}, b_w) : \|\tilde{w}\| \in [1/C_1, C_1], |b_w| \leq C_2\}$, and consider empirical gradient descent iterates: $w_{t+1} = w_t - \eta\nabla\widehat{L}(w_t)$. For a suitable constant learning rate η, when starting from $w_0 = (\tilde{w}_0, 0)$ where $\tilde{w}_0$ is randomly initialized from a radially symmetric distribution, when given $\mathrm{poly}(d, 1/\varepsilon, B_Y)$ i.i.d. samples from the data distribution D, with at least constant probability $c_3' > 0$ one of the iterates w_T of gradient descent after poly$(d, \frac{1}{\varepsilon})$ steps satisfies $L(w_T) = O(OPT) + 2\varepsilon$.*

We remark that the above theorem also holds under our weaker distributional assumptions in Section B with an additional sub-Gaussianity assumption on $\widetilde{\mathcal{D}}_x$, as evident from the proof that follows. In order to prove Theorem D.1, we first introduce the following definitions and lemmas. The following definition is a standard tool for establishing uniform convergence guarantees and is deeply related to the notion of Rademacher Complexity. For further details please refer to Shalev-Shwartz & Ben-David (2014).

Definition D.2 (Representativeness)**.** Given data samples $S = \{z_1, ..., z_n\} \in \mathcal{Z}^n$ and a function class $\mathcal{F} = \{f : \mathcal{Z} \to \mathbb{R}\}$, the representativeness of S with respect to $\mathcal{F}$ is

$$\mathrm{Rep}(\mathcal{F}, S) = \sup_{f\in\mathcal{F}} \mathbb{E}[f(z)] - \frac{1}{n}\sum_{i=1}^{n} f(z_i)$$

Note that representativeness is a random variable. The following lemma quantifies the convergence property of representativeness with respect to the loss function gradient through analyzing its Rademacher complexity.

Lemma D.3 (Concentration of Representativeness)**.** *For absolute constants $c_1, c_2, c_3 > 0$, with probability at least $1 - \kappa$, the representativeness of random samples $S = \{(\widetilde{x}_i, y_i)\}_{i=1}^n \sim_{i.i.d.} D$ with respect to the function class $\mathcal{F}_j = \{(\sigma(w^\top x) - y)\sigma'(w^\top x)x_j \ : \ \|w\| \leq C_1\}$, $\mathrm{Rep}(\mathcal{F}_j, S)$ is bounded by*

$$\mathrm{Rep}(\mathcal{F}_j, S) \leq \frac{c_1 d B_Y C\sqrt{d\log(Cn)}}{\sqrt{n}} + \sqrt{\frac{c_3 d B_Y^2 \log(4/\kappa)}{n}}$$

where for all y_i, $|y_i| \leq B_Y$.

Proof. Note that $\mathbb{E}[\mathrm{Rep}(\mathcal{F}_j, S)] \leq 2\,\mathbb{E}[R(\mathcal{F}_j \circ S)]$, where $R(\mathcal{F}_j \circ S)$ is the Rademacher Complexity of the set $\{\{(\sigma(w^\top x_i) - y_i)\sigma'(w^\top x_i)x_{ij}\}_{i=1}^n : \|w\| \leq C_1\}$ (Lemma 26.2 of Shalev-Shwartz &

Ben-David (2014)). Hence, combining it with McDiarmid's inequality for almost-bounded difference functions (see Kutin (2002)), with probability at least $1-\kappa$ we get

$$\mathrm{Rep}(\mathcal{F}_j, S) \leq \mathbb{E}[\mathrm{Rep}(\mathcal{F}_j, S)] + \sqrt{\frac{c_3 d B_Y^2}{n} \log\left(\frac{4}{\kappa}\right)} \leq 2\,\mathbb{E}[R(\mathcal{F}_j \circ S)] + \sqrt{\frac{c_3 d B_Y^2}{n} \log\left(\frac{4}{\kappa}\right)}$$

For the first term, by definition we have

$$R(\mathcal{F}_j \circ S) = \frac{1}{n}\mathop{\mathbb{E}}_{s}\Big[\sup_{\|w\|\leq C_1} \sum_{i=1}^{n} s_i x_{ij} \sigma'(w^\top x_i)(w^\top x_i - y_i)\Big]$$

where $\{s_i\}_{i=1}^n$ are i.i.d. Rademacher random variables. Given the sample S, let R_S be the maximum ℓ_2 norm of a vector $x_i \in S$. We can further upper bound the above as

$$R(\mathcal{F}_j \circ S) \leq \frac{1}{n}\mathop{\mathbb{E}}_{s}\Big[\sup_{\|w\|\leq C_1} \sum_{i=1}^{n} s_i x_{ij} \sigma'(w^\top x_i)(w^\top x_i)\Big] + \frac{1}{n}\mathop{\mathbb{E}}_{s}\Big[\sup_{\|w\|\leq C_1} \sum_{i=1}^{n} s_i x_{ij} \sigma'(w^\top x_i) y_i\Big]$$

For the second term above we can use Massart's finite class lemma Shalev-Shwartz & Ben-David (2014) and noticing that the sup is only over $O(n^{d+1})$ different hypotheses (since only sign of $w^\top x_i$ matters, and we can use Sauer-Shelah's lemma with linear classifiers in d dimensions Shalev-Shwartz & Ben-David (2014)), we get that

$$\frac{1}{n}\mathop{\mathbb{E}}_{s}\Big[\sup_{\|w\|\leq C_1} \sum_{i=1}^{n} s_i x_{ij} \sigma'(w^\top x_i) y_i\Big] \leq O(\frac{1}{\sqrt{n}} R_S B_Y \sqrt{d \log n}).$$

To bound the first term, for an appropriate ε to be chosen later, let H_ε be a minimal ε-cover for the set $\{w \in \mathbb{R}^d : \|w\| \leq C\}$. It is well known that $|H_\varepsilon| = O(C/\varepsilon)^d$ Shalev-Shwartz & Ben-David (2014). For any $w \in \mathbb{R}^d$ such that $\|w\| \leq C$ we will denote by w_ε the closest vector to w (in ℓ_2 distance) in the set H_ε. Then we can write

$$\begin{aligned} R(\mathcal{F}_j \circ S) \leq{}& \frac{1}{n}\mathop{\mathbb{E}}_{s}\Big[\sup_{w\in H_\varepsilon} \sum_{i=1}^{n} s_i x_{ij} \sigma'(w^\top x_i)(w^\top x_i)\Big] \\ &+ \frac{1}{n}\mathop{\mathbb{E}}_{s}\Big[\sup_{\|w\|\leq C_1} \Big(\sum_{i=1}^{n} s_i x_{ij} \sigma'(w^\top x_i)(w^\top x_i) - \sum_{i=1}^{n} s_i x_{ij} \sigma'(w_\varepsilon^\top x_i)(w_\varepsilon^\top x_i)\Big)\Big] \end{aligned}$$

Noticing that $|(w^\top - w_\varepsilon^\top)\cdot x_i| \leq \varepsilon R_S$, and the fact that $|\sigma'(t_1)t_1 - \sigma'(t_2)t_2| \leq |t_1 - t_2|$, we get that

$$R(\mathcal{F}_j \circ S) \leq \frac{1}{n}\mathop{\mathbb{E}}_{s}\Big[\sup_{w\in H_\varepsilon} \sum_{i=1}^{n} s_i x_{ij} \sigma'(w^\top x_i)(w^\top x_i)\Big] + O(\varepsilon R_S B_Y) + O(\frac{1}{\sqrt{n}} R_S B_Y \sqrt{d \log n}). \tag{35}$$

For the first term above we apply Massart's finite class lemma Shalev-Shwartz & Ben-David (2014) to get that

$$\frac{1}{n}\mathop{\mathbb{E}}_{s}\Big[\sup_{w\in H_\varepsilon} \sum_{i=1}^{n} s_i x_{ij} \sigma'(w^\top x_i)(w^\top x_i)\Big] \leq O(\frac{1}{\sqrt{n}} R_S C \sqrt{\log(|H_\varepsilon|)}). \tag{36}$$

From (35) and (36) we get that

$$R(\mathcal{F}_j \circ S) = O(\frac{1}{\sqrt{n}} R_S B_Y C \sqrt{d \log(Cn/\varepsilon)} + \varepsilon R_S B_Y).$$

Substituting $\varepsilon = 1/\sqrt{n}$ above we get that

$$R(\mathcal{F}_j \circ S) = O\big(\frac{1}{\sqrt{n}} R_S B_Y C \sqrt{d \log(Cn)}\big).$$

Finally, taking the expectation over S and using standard property of Gaussians we get that

$$\mathbb{E}[R(\mathcal{F}_j \circ S)] = O\big(\frac{1}{\sqrt{n}} dB_Y C\sqrt{d\log(Cn)}\big).$$

This completes the proof. □

With these lemmas, we are now ready to prove Theorem D.1.

Proof of Theorem D.1. The proof consists of three parts highly identical to that of Theorem 1.1, hence we only highlight the main difference. As in the proof of Theorem 1.1 we will assume that $\|v\|_2 = 1$; note that this is without loss of generality from Proposition C.1. Also as in the proof of Theorem 1.1, we will only argue at one of the iterates T satisfies the required guarantee (this may not be the last iterate).

In the first part, we rewrite the update rule as

$$w_{t+1} = w_t - \eta\nabla L(w_t) + \eta\Big(\nabla L(w_t) - \nabla\widehat{L}(w_t)\Big) = w_t - \eta\nabla L(w_t) + \eta\zeta_t,$$

$$\text{where } \zeta_t \coloneqq \nabla L(w_t) - \nabla\widehat{L}(w_t), \text{ and } g(\zeta_t) \coloneqq -\eta\langle\nabla L(w_t), \zeta_t\rangle + \langle w_t - v, \zeta_t\rangle + \frac{\eta\|\zeta_t\|^2}{2}. \quad (37)$$

We obtain the improvement in each iteration as

$$\|w_t - v\|^2 - \|w_{t+1} - v\|^2 = 2\eta\langle\nabla L(w_t), w_t - v\rangle - \eta^2\|\nabla L(w_t)\|^2 - 2\eta g(\zeta_t)$$

Note that $2\eta g(\zeta_t)$ is a random variable that depends on z and can possibly be negative. We will later use a uniform convergence bound in Lemma D.3 to bound both $\|\zeta_t\|$ and hence $|g(\zeta_t)|$ with high probability for all w that is bounded by a fixed constant. Conditioned on this high probability event (given in Lemma D.3), the rest of the analysis is deterministic. Recall that $\varepsilon > 0$ is the parameter denoting the desired error. We will maintain the invariants that when gradient descent is still in progress (or we haven't encountered a time step with our desired guarantees), $|g(\zeta_t)| \le \varepsilon$, and $\|w_t - v\|$ is bounded by a constant.

Recall that

$$g(\zeta_t) = -\eta\langle\nabla L(w_t), \zeta_t\rangle + \langle w_t - v, \zeta_t\rangle + \frac{\eta\|\zeta_t\|^2}{2}$$

By applying the upper bound for $\|\nabla L(w_t)\|$ as in the population argument (see Equation 20 in the proof of Theorem 1.1), we get for some constant $C' > 0$

$$\begin{aligned}|g(\zeta_t)| &\le \eta\|\nabla L(w_t)\|\|\zeta_t\| + \|w_t - v\|\|\zeta_t\| + \frac{\eta\|\zeta_t\|^2}{2}\\ &\le \eta\sqrt{C'd(\|w_t - v\|^2 + OPT)}\|\zeta_t\| + \|w_t - v\|\|\zeta_t\| + \frac{\eta\|\zeta_t\|^2}{2}\end{aligned}$$

In addition, since at this point gradient descent is still running, $C'\sqrt{OPT} \le \|w_t - v\|$, hence with suitable constant $C'' > 0$ we can further write

$$|g(\zeta_t)| \le \eta\sqrt{C_G}dC''\|w_t - v\|\|\zeta_t\| + \frac{\eta\|\zeta_t\|^2}{2}$$

Again, while gradient descent is still in progress, our induction argument establishes that $\|w_t - v\|^2 - \|w_{t+1} - v\|^2$ is lower-bounded by a non-negative amount, hence $\|w_{t+1} - v\| \le \|w_t - v\| \le \ldots \le \|w_0 - v\| = O(1)$ which also establishes that every $\|w_t\|$ is upper-bound by a constant. Let T be the time step until which all of the above properties hold (otherwise we have already encountered an iterate where we get the required guarantee). Therefore we can conclude that $|g(\zeta_t)| \le O(\|\zeta_t\|) + \frac{\eta\|\zeta_t\|^2}{2}$, $\forall t \le T$.

We now proceed to bound the magnitude of $\|\zeta_t\|$. Using Lemma D.3, for each coordinate of ζ_t we sample $\mathrm{poly}(d, 1/\varepsilon, B_Y)$ data points so that with probability $1 - \kappa/(d+1)$ its magnitude is at most

$\varepsilon/(d+1)$. We then take the union bound over all $d+1$ coordinates and set $\kappa = 1/d^3$ to conclude that with high probability,

$$\forall t \leq T, \;\; \|g(\zeta_t)\| \leq \varepsilon, \text{ and } \|\zeta_t\|_2 \leq \varepsilon/c. \tag{38}$$

Now, since we have showed that $\|g(\zeta_t)\|$ remains bounded by ε, identical to our argument in the proof for Theorem 1.1 except modifying the induction hypothesis to be $\|w_t - v\| > C_p \gamma^{-1/2}\sqrt{OPT + 2\varepsilon}$, we conclude that after $T \leq O(d(OPT + 2\varepsilon)^{-1})$ iterations we get $\|w_t - v\|^2 \leq O(OPT) + 2\varepsilon$, and similarly by Lemma 4.4 this implies both $F(w_T)$ and $L(w_T)$ are at most $O(OPT) + 2\varepsilon$.

Proceeding to the second part of the proof, we will show that while gradient descent is still running, $F(w_t)$ continues to decrease. We rewrite the expression given in Lemma A.2 as

$$F(w_t - \eta \nabla \widehat{L}(w_t)) \leq F(w_t) - \eta\langle \nabla F(w_t), \nabla L(w_t)\rangle + \ell\eta^2 \|\nabla L(w_t)\|^2 - \eta\langle \nabla F(w_t), \zeta_t\rangle + \ell\eta^2\|\zeta_t\|^2$$

At this point, note that we can still argue that $\|\nabla F(w_t)\| > C_G\sqrt{OPT + 2\varepsilon}$ directly by Lemma 4.3, for some constant $C''' > 0$, we can hence upper-bound the second and third terms by directly applying Equation 21, yielding

$$\eta\Big(- C'''(OPT + 2\varepsilon) + \ell\eta C_L d\|w_t - v\|^2\Big) - \eta\langle \nabla F(w_t), \zeta_t\rangle + \ell\eta^2\|\zeta_t\|^2$$

$$\leq \eta\Big(- C'''(OPT + 2\varepsilon) + \ell\eta C_L d\|w_t - v\|^2 + C_F d\|w_t - v\|\|\zeta_t\| + \ell\eta\|\zeta_t\|^2\Big).$$

Therefore by applying the same analysis as in the population case, and using (38) we have that the above upper bound is

$$\leq \eta\Big(- C'''(OPT + 2\varepsilon) + \ell\eta C_L d\|w_t - v\|^2 + \varepsilon/c\Big) \leq 0$$

Hence $F(w_t)$ continues to decrease, hence $F(w_t) \leq F(0) - \delta$.

Finally, by Lemma B.3 with constant probability gradient descent starts at a point such that $F(w_0) \leq F(0) - \delta$, hence the proof follows.